(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2004 (29.12.2004)

PCT

(10) International Publication Number
WO 2004/114081 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number:
PCT/US2004/019471

(22) International Filing Date: 18 June 2004 (18.06.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/480,038 20 June 2003 (20.06.2003) US

(71) Applicant *(for all designated States except US)*: **PARADIGM GENETICS, INC.** [US/US]; 108 Alexander Drive, Post Office Box 14528, Research Triangle Park, NC 27709-4528 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **COFFIN, Marie** [US/US]; 102 Eagle Court, Cary, NC 27511 (US). **LAWRENCE, Matthew** [US/US]; 300 Brown Circle, Rolesville, NC 27571 (US).

(74) Agent: **KIEFER, Laura, A.**; Paradigm Genetics, Inc., 108 Alexander Drive, Post Office Box 14528, Research Triangle Park, NC 27709-4528 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2004/114081 A2

(54) Title: METHODS AND SYSTEMS FOR CREATION OF A COHERENCE DATABASE

(57) Abstract: The present invention provides methods and systems for organizing complex biological data in a database schema that facilitates data analysis in a biological context. Specifically, the methods and systems of the present invention pertain to the creation of an integrated relational database schema for recording and organizing summary data from experiments, relating data from disparate data streams, and relating data to reference information sources. The invention is useful in multiple applications, including applications in the agricultural, pharmaceutical, forensic, biotechnology, and nutriceutical industries.

streams in an integrated relational database schema that allows relating of empirical data to reference information sources, and facilitates recognition and identification of trends and relationships within complex data. Methods and systems of the present invention are useful in creating a coherence database comprised of at least one data table containing a unique identifier of at least one experiment; at least one data table containing a unique identifier of at least one biological sample; at least one data table containing data measurements from the biological sample; at least one data table containing attribute information; placement of all of the data tables in an integrated relational database schema; and relating the data tables of the integrated relational database schema, through the attribute information, to at least one reference information source. The integrated relational database schema resulting from the methods and systems of the present invention allows data to be examined within a biological context.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 depicts the flow of information in an exemplary coherence database schema.

Figure 2 depicts the schema of the coherence database (104) of Figure 1 and is described in detail in the Specific Examples that follow.

DETAILED DESCRIPTION OF THE INVENTION

Definitions:

Identifying a "baseline" or control value is essential to biological experimentation and provides, but is not limited to, a mechanism for distinguishing perturbed from unperturbed. A baseline is used in the invention to standardize data to a common or commonly relevant unit of measure. The term "baseline" is herein used to refer to and is interchangeable with "reference" and "control." Baseline populations consist, for example, of data from organisms of a particular group, such as healthy or normal organisms, or organisms diagnosed as having a particular disease state, pathophysiological condition, or other physiological state of interest. An example of the use of a baseline is the expression of data measurements as standard deviations from the corresponding baseline mean.

Polymeric compounds, such as glycogen, are important participants in metabolic reactions as a source of metabolites, but are not chemically defineable (i.e. an input/output to metabolism). Thus, polymeric compounds are excluded from the definition of metabolite as used herein.

Metabolites of xenobiotics (chemical compounds foreign to the body or to living organisms) are neither native, required for maintenance or growth, nor required for normal function of a cell, and thus are not metabolites as used herein. However, it is useful to monitor xenobiotics when observing the effects of a drug therapy program, or in experimentally determining the effects of a compound on an individual. Essential or nutritionally required compounds are not synthesized *de novo,* (*i.e.* not native), but are required for the maintenance, growth, or normal function of a cell. Therefore, essential or nutritionally required compounds are metabolites as defined herein.

"Morphology" refers to the form and structure of an organism or any of its parts. Morphology is one way of referring to a phenotype.

"Peak" refers to the readout from any type of spectral analysis or metabolite analysis instrumentation, as is standard in the art, and can represent one or more chemical components. The instrumentation can include, but is not limited to, liquid chromatography (LC), high-pressure liquid chromatography (HPLC), mass spectrometry (MS), hyphenated detection systems such as MS-MS or MS-MS-MS, gas chromatography (GC), liquid chromatography/mass spectroscopy (LC-MS), gas chromatography/mass spectroscopy (GC-MS), Fourier transform-ion cyclotron resonance-mass spectrometry (FT-MS), nuclear magnetic resonance (NMR), magnetic resonance imaging (MRI), Fourier Transform InfraRed (FT-IR), and inductively coupled plasma mass spectrometry (ICP-MS). It is further understood that mass spectrometry techniques include, but are not limited to, the use of magnetic-sector and double focusing instruments, transmission quadrapole instruments, quadrupole ion-trap instruments, time-of-flight instruments (TOF), Fourier transform ion cyclotron resonance instruments (FT-MS), and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). It is understood that the phrase "mass spectrometry" is used interchangeably with "mass spectroscopy" in this application.

herein is a support tool that enables other applications or software tools to be most successfully applied in data analysis, and the invention presented herein facilitates recognition and identification of trends and relationships within complex data.

Accordingly, the present invention provides methods and systems for recording and organizing summary data from experiments, relating data from disparate data streams, and relating data to reference information sources. The methods and systems of the present invention are useful in numerous applications, such as determining gene function; identifying and validating drug and pesticide targets; identifying and validating drug and pesticide candidate compounds; profiling of drug and pesticide compounds; predicting the toxicological impact of a drug or pesticide compound; producing a compilation of health or wellness profiles; identifying suites of compounds, proteins, genes, or combinations thereof to act as biomarkers of a biological status; determining compound sites of action; identifying unknown samples; and numerous other applications in the agricultural, pharmaceutical, nutraceutical, forensic, and biotechnology industries.

Thus, in one embodiment, the coherence database resulting from the methods and systems of the present invention is comprised of at least one data table containing a unique identifier of at least one experiment; at least one data table containing a unique identifier of at least one biological sample; at least one data table containing summary data measurements from the biological sample; at least one data table containing information about attributes pertaining to the summary data measurements; placing the data tables in an integrated relational database schema; and relating the data tables of the integrated relational database schema, through the attribute information, to at least one reference information source. The terms "data table" and "table" are used interchangably in the present application.

Experimental design and conditions include any factors that can be used to stratify data. The experimental design and conditions recorded may include, but are not limited to, organism species; organism type within a species (such as sex (male or female); age; race; body type (obese, thin, tall, short); behaviors such as smoking or exercising; presence or absence of disease; mutant type; or other factors contributing to a patient profile); sample type (tissue or fluids such as blood or urine); treatment type (drug or pesticide compound, mode of administration, length of time administered and amount

technologies can be used to obtain the same type of gene information, including high-density array spotting on glass or membranes and quantitative reverse transcription and PCR.

Phenotype refers to observable physical or biochemical/metabolic characteristics of an organism, as determined by genetic and environmental factors. For example, in an *Arabidopsis thaliana* plant model system, a phenotype can be described by using distinctly defined attributes such as, but not limited to, number of: abnormal seeds, cotyledons, normal seeds, open flowers, pistils per flower, senescent flowers, sepals per flower, siliques, and stamens. Perturbation of a biological system is often indicated by a phenotypic trait. In humans, a perturbed biological system may result in symptoms of disease such as chest pain, signs such as elevated blood pressure, or observable physical traits such as those exhibited by individuals afflicted with Trisomy 21. A normal phenotype is useful as a baseline value against which a physiological status can be measured.

Medical history, examination, and testing techniques are well known to medical practitioners and data derived from the same can be used in practicing the methods and systems of the present invention. For example, in cases where a practitioner is examining a patient to determine the likelihood, existence, or extent of coronary heart disease (CHD), phenotypic traits observed or identified in a clinical setting include, but are not limited to, risk factors such as blood pressure, cigarette smoking, total cholesterol (TC), low density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), and diabetes. P.G. McGovern et al., 334 NEW ENG. J. MED. 884-890 (1996). Additonal phenotypic characteristics such as body weight, family history of CHD, hormone replacement therapy, and left ventricular hypertrophy are also useful in determining CHD risk. It is common in the medical arts to scale or score a patient's condition based on a set of phenotypic signs and symptoms. For example, predictive models have been described based on blood pressure, cholesterol, and LDL-C categories as identified by the National Cholesterol Education Program and the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure. P.W.F. Wilson et al., 97 CIRCULATION 1837-1847 (1998) (incorporated herein by reference). Furthermore, predictive outcome models have also been described for patients

Syndrome International Prognostic Scoring System; Nonbiliary Cirrhosis Prognostic Criteria for One Year Survival; Obesity Management Guidelines (National Institutes of Health/NHLBI); Perioperative Cardiac Evaluation (NHLBI); Polycythemia Vera Diagnostic Criteria; Prostatism Symptom Score; Ranson Criteria for Acute Pancreatitis; Renal Artery Stenosis Prediction Rule; Rheumatoid Arthritis Criteria (American Rheumatism Association); Romhilt-Estes Criteria for Left Ventricular Hypertrophy; Smoking Cessation and Intervention (NHLBI); Sore Throat (Pharyngitis) Evaluation and Treatment Criteria; Suggested Management of Patients with Raised Lipid Levels (NHLBI); Systemic Lupus Erythematosis American Rheumatism Association 11 Criteria; Thyroid Disease Screening for Females More Than 50 Years Old (NHLBI); and Vector and Scalar Electrocardiography.

Still other phenotypic traits could be observed or identified by x-ray; cardiac and vascular angiography; electrocardiogaphy; blood pressure (BP) examination; pulse; weight and height; ideal body weight or BMI; retinal examination; thyroid examination; carotid bruits; neck vein examination; congestive heart failure (CHF) signs; palpable intercostal pulses; cardiovascular examination traits including, but not limited to, S4 gallop, tachycardia, bradycardia, heart sounds, aortic insufficiency, murmur, and echocardiography; abdominal examination; genitourinary examination; peripheral vascular disease examination; neurologic examination; and skin examination. In addition to standard x-ray technologies, numerous imaging technigues are also useful in observing and identifying phenotypic traits including, but not limited to, ultrasound, magnetic resonance imaging (MRI), positron emission tomography (PET), single photon emission computed tomography (SPECT), x-ray transmission, x-ray computed tomography (X-ray CT), ultrasound electrical impedance tomography (EIT), electrical source imaging (ESI), magnetic source imaging, (MSI) laser optical imaging.

Metabolite or biochemical analysis (also referred to as biochemical profiling or BCP) refers to an analysis of organic, inorganic, and/or bio-molecules (hereinafter collectively referred to as "small molecules") of a cell, cell organelle, tissue and/or organism. It is understood that a small molecule is also referred to as a metabolite. Techniques and methods of the present invention employed to separate and identify small molecules, or metabolites, include but are not limited to: liquid chromatography (LC),

are undetectable to the human eye. One example of tissue feature analysis is described in Kriete et al., 4 Genome Biology R32.1-.9 (2003).

Attributes refer to any information useful in accessing or querying data, and may include, but are not limited to, information such as compound molecular weight, compound structure, gene sequence, gene annotation, gene splice variants, genes encoding particular proteins, protein molecular weight, protein isoelectric point, protein active domain sequence and/or consensus sequence, annotation and/or references pertaining to phenotypic or morphological data, tissue type, treatment type, and mutant type. Attributes are useful in relating empirical data to reference information sources.

Reference information sources include, but are not limited to, KEGG (Kyoto Encyclopedia of Genes and Genomes, Institute for Chemical Research, Kyoto University, Japan), BRENDA (The Comprehensive Enzyme Information System, Institute of Biochemistry, University of Cologne, Germany), Expert Protein Analysis System (ExPASy), or any other information source that provides a biological context for data analysis, including a proprietary data source. The biological context may include a biochemical pathways context, which may include substrates, products, and enzymes (all metabolites) and the genes that encode the metabolites. In another embodiment, a signal transduction context or a protein-binding (protein-protein interactions) context, such as cell surface binding, protein kinase reactions (signal transduction), cytokine binding (signal transduction), or antibody binding, is provided. In another embodiment, a cellular organelle context, such as a mitochondrial context, a cellular context, a tissue context, an organ context, an organ system context, or an entire organism context, is provided. In another embodiment, a chromosomal context, such as genes or metabolites represented on a chromosome map of a particular organism, is provided. In another embodiment, an image context is provided, such as a CAT (or CT) scan, an MRI, a histology image such as a section of an organ or tissue, a depiction of a human body, a depiction of a human tissue, organ, or organ system, a depiction of a leaf, a root, a stem, a flower, a seed, an entire plant, or any image of an organism or any part thereof. In yet another embodiment, a protein structure or model context is provided, such as the structure of an enzyme complex, on which genes are superimposed. In another embodiment, a context of global architecture of genetic interactions on protein networks is provided (O. Ozier et al., 21

biological sample are comprised of a first data type in a first data table, a second data type in a second data table, and a third data type in a third data table.

In one embodiment of the present invention the data measurements include RNA data (gene expression profiling analysis), phenotypic data, and metabolite data (biochemical profiling analysis), but one skilled in the art will understand that data from any technology or process may be utilized in the methods and systems of the invention. Further, it is understood by one skilled in the art that data from any biological organism (alive or dead) or part thereof may be incorporated into a coherence database. Suitable biological organisms include, but are not limited to, plants, such as *Arabidopsis* (*Arabidopsis thaliana*), corn and rice, fungal organisms including *Magnaporthe grisea, Saccharomyces cerevisiae*, and *Candida albicans*, and mammals, including rodents, rabbits, canines, felines, bovines, equines, porcines, and human and non-human primates.

Figure 1 depicts the flow of information in an exemplary coherence database schema. Information about experiments (101) represents detailed information pertaining to experimental design and conditions relating to the experimental design. In one instance, information about experiments (101) may be recorded in a laboratory information management system (LIMS). Each experiment is assigned a unique identifier. Unique experiment identifiers recorded in the coherence database (104) are related to detailed experimental information (101). Experiment information found in a coherence database includes a single unique identifier for an entire experiment, and attributes, which are specific references to particular features of the experiment. Experimental data (102) represents raw unmanipulated experimental data acquired directly from scientific instrumentation. The experimental data may be subject to processes such as quality control and quality assurance procedures. A statistical processor (103), in which the experimental data (102) is processed into summary data, is related to information about experiments (101). External data source I (105), external data source II (106), and proprietary data source (107) represent reference information sources external to the coherence database and separate from empirical information (experimental design (101) and experimental data (102)). Such separation of empirical data and reference data allows security measures to be implemented for protecting empirical data without hampering access to referenc information sources. External data

table in a second database. In another example, GEP data are recorded in a first data table and phenotypic data are recorded in a second data table, both of which are recorded in a first database, and BCP data are recorded in a third data table in a second database. In another example, BCP data are recorded in a first data table and phenotypic data are recorded in a second data table, both of which are recorded in a first database, and GEP data are recorded in a third data table in a second database.

In another embodiment, the coherence database resulting from the methods and systems of the present invention is comprised of at least one data table containing a unique identifier of at least one experiment; at least one data table containing a unique identifier of at least one biological sample; at least one data table containing summary data measurements from the biological sample, wherein the summary data measurements are from genes, proteins, metabolic compounds, or phenotype (including morphology or histology); at least one data table containing information about attributes pertaining to the summary data measurements; placing all of the data tables in an integrated relational database schema; and relating the data tables of the integrated relational database schema, through the attribute information, to at least one reference information source.

In a further embodiment, the coherence database resulting from the methods and systems of the present invention is comprised of at least one data table containing a unique identifier of at least one experiment; at least one data table containing a unique identifier of at least one biological sample; at least one data table containing summary data measurements from the biological sample; at least one data table containing information about attributes pertaining to the summary data measurements; placing the data tables in an integrated relational database schema; and relating the data tables of the integrated relational database schema, through the attribute information, to at least one reference information source, wherein the at least one reference information source is KEGG and/or BRENDA and/or ExPASy and/or any biochemical pathway or network information source.

In still another embodiment, the coherence database resulting from the methods and systems of the present invention is comprised of at least one data table containing a unique identifier of at least one experiment; at least one data table containing a unique identifier of at least one biological sample; at least one data table containing summary

isoelectric point, tissue type, treatment type, mutant type, and/or phenotype/morphology annotation and references to publications; placing the data tables in an integrated relational database schema; and relating the data tables of the integrated relational database schema, through the attribute information, to at least one reference information source, wherein the reference information source is KEGG and/or BRENDA and/or ExPASy and/or any biochemical pathway or network information source. It is understood by those of ordinary skill in the art that not all possible examples of integrated relational database schema are listed here and, accordingly, additional ways of creating a coherence database fall under the scope of the present invention.

EXPERIMENTAL

Example 1

Tables of Experimental Design and Conditions

Figure 2 portrays a detailed coherence database schema, with the contents of each data table specified. In the current example, experimental design and conditions using *Arabidopsis thaliana* plants were determined and the data tables were populated accordingly. Referring to Figure 2, table 221 represents a data table containing details about sample (tissue) type, table 222 represents a data table containing details about organism mutant type (for example, a transgenic organism), table 223 represents a data table containing details about the experimental treatment type, and table 224 represents a data table containing details about the organism species type. These four data tables are related in various ways to summary data tables populated with information of various types regarding the experiments. Referring still to Figure 2, the tissue type (221) and species type (224) are related to the AT Line summary set data table (216). The "AT Line" refers to the *Arabidopsis thaliana* plant line and contains details of the specifics of the plant line, including genetic information. Table 215 is a look-up data table providing a workflow tracking mechanism for the large number of plants processed and is related to AT Line summary set data table (216). Tissue type (221), species type (224), mutant type (222) and treatment type (223) are also related to the treatment summary set data table (217), the time summary set data table (218), the tissue summary set data table

Genomic Research) and GenBank databases. Table 207 is related to the gene data summary set data table (208). Table 206 is a look-up data table providing information (including nucleotide sequence information) directed to different genes or gene fragments used in the gene expression profiling studies, and is related to table 207. Table 210 contains attributes pertaining to biochemical compounds or metabolites, such as compound name, chemical formula, CAS number, and KEGG compound identifier. Table 210 is related to the biochemical profiling summary data set data table (211). Attributes are useful in accessing or querying data in the coherence database, and are used to relate data in the coherence database to external information sources.

Example 4

Primary Summary Set Table

As is shown in Figure 2, a central data table called the summary set data table (209) was related to a look-up data table containing descriptions of the summary set types (204). Table 204 contains summary set types such as mutant type, treatment type, time, tissue type, and *Arabidopsis* line. The primary summary set data table (209) contains information from throughout the coherence database, allowing queries of any of the data contained therein.

Example 5

Acetaminophen Activity in Rat Tissue

Acetaminophen overdose is one of the leading causes of liver failure. In this experiment, rats were dosed with acetaminophen and livers were harvested across a time course. Two doses of acetaminophen were used (50 mg/kg and 1500 mg/kg), as well as a control group that received no acetaminophen. The harvest times were 6, 18, 24, and 48 hours. Three rats (biological replicates) were in each treatment group, wherein a treatment group is defined as each combination of dose and time. Referring now to Figure 2, experimental information was entered into data tables 221, 223, and 224 (tissue_type = liver, treatment_type = acetaminophen, treatment_concentration = dose, and species_type = rat) in the coherence database and is also summarized in table 217 (treatment_summary_set), thus allowing comparison of two or more treatment types.

At this point, scientists queried the database and discovered that more compounds were perturbed at the 18 hour timepoint than any other. Consequently, a pathway query tool was used to obtain a list of pathways showing metabolic perturbation at the 18 hour timepoint. Using a pathway viewing tool on the 18 hour timepoint data led to the conclusion that the nitrogen metabolism pathway was most likely the source of the primary metabolic disturbance. This exemplifies how the coherence database of the present invention facilitated data analysis by enabling queries using aspects of the experimental design and by using attributes to relate to a data source (KEGG) external to the coherence database.

Example 6

Herbicide Mode of Action Experiment

Eighteen known herbicides were used to treat *Arabidopsis* plants. The first experiment was a dose-response experiment, used to determine the Minimum Inhibitory Concentration (MIC) and Time to reach complete inhibition (TMIC) for each herbicide. Following this preliminary work, an experiment was performed in which *Arabidopsis* plants were treated with the 18 herbicides. For each herbicide, the MIC was used, and plants were harvested at 30%, 50% and 70% of TMIC timepoints. Because the timepoints were different for herbicides that act at different rates, matched control plants were harvested at the same timepoints. Before harvesting, each plant was rated on 12 phenotypic measurements determined to be relevant to herbicide action. From the leaf tissue samples, biochemical profiling (as in Example 5), and gene expresssion profiling (GEP) were carried out.

The standardized differences from matched controls were calculated as described in Example 5, using the biochemical profiling data, gene expression profiling data, and the phenotypic data. Referring now to Figure 2, a summary set was created for each herbicide at each timepoint, the herbicide (treatment) name and timepoint for each summary set were recorded in the treatment_summary_set data table (217), and the summary set description was recorded in the summary_set data table (209). The standardized differences from matched controls were recorded in the bcp_summary (211), gep_summary (208), and pheno_summary data tables (203). The compounds in

Published references and patent publications cited herein are incorporated by reference as if terms incorporating the same were provided upon each occurrence of the individual reference or patent document. While the foregoing describes certain embodiments of the invention, it will be understood by those skilled in the art that variations and modifications may be made that will fall within the scope of the invention. The foregoing examples are intended to exemplify various specific embodiments of the invention and do not limit its scope in any manner.

6. The method of claim 1, wherein the summary data measurements are comprised of phenotypic data measurements and gene expression profiling data measurements.

7. The method of claim 1, wherein the summary data measurements are comprised of phenotypic data measurements and biochemical profiling data measurements.

8. The method of claim 1, wherein the summary data measurements are comprised of gene expression profiling data measurements, phenotypic data measurements, and biochemical profiling data measurements.

9. The method of claim 1, wherein the at least one reference information source is selected from the group consisting of KEGG, ExPASy or Brenda.

10. A method for creating a database, comprising:
    a) creating at least one data table containing a unique identifier of at least one experiment;
    b) creating at least one data table containing a unique identifier for at least one biological sample obtained from the experiment of step (a);
    c) creating at least two data tables containing summary data measurements from said at least one biological sample;
    d) creating at least one data table containing information about attributes pertaining to the summary data measurements of step (c);
    e) placing the data tables in steps (a) through (d) in an integrated relational database schema; and
    f) relating the data tables in the integrated relational database schema to at least one reference information source, wherein the attributes of step (d) provide the relationship between the integrated relational database schema data and the at least one reference information source.

b) means for creating at least one data table containing a unique identifier for at least one biological sample obtained under the experiment of step (a);

c) means for creating at least one data table containing summary data measurements from said at least one biological sample;

d) means for creating at least one data table containing information about attributes pertaining to the summary data measurements of step (c);

e) means for placing the data tables in steps (a) through (d) in an integrated relational database schema; and

f) means for relating the data tables in the integrated relational database schema to at least one reference information source, wherein the attributes of step (d) provide the relationship between the integrated relational database schema data and the at least one reference information source.

19. The system of claim 18, wherein the summary data measurements are comprised of gene expression profiling data measurements.

20. The system of claim 18, wherein the summary data measurements are comprised of biochemical profiling data measurements.

21. The system of claim 18, wherein the summary data measurements are comprised of gene expression profiling data measurements and biochemical profiling data measurements.

22. The system of claim 18, wherein the summary data measurements are comprised of phenotypic data measurements.

23. The system of claim 18, wherein the summary data measurements are comprised of phenotypic data measurements and gene expression profiling data measurements.

28. The system of claim 27, wherein the summary data measurements of step (c) are comprised of a first data type in a first data table and a second data type in a second data table.

29. The system of claim 27, wherein the summary data measurements of step (c) are comprised of a first data type in a first data table, a second data type in a second data table, and a third data type in a third data table.

30. The system of claim 27, wherein the summary data measurements are comprised of gene expression profiling data measurements and biochemical profiling data measurements.

31. The system of claim 27, wherein the summary data measurements are comprised of phenotypic data measurements and gene expression profiling data measurements.

32. The system of claim 27, wherein the summary data measurements are comprised of phenotypic data measurements and biochemical profiling data measurements.

33. The system of claim 27, wherein the summary data measurements are comprised of gene expression profiling data measurements, phenotypic data measurements, and biochemical profiling data measurements.

34. The system of claim 27, wherein the at least one reference information source is selected from the group consisting of KEGG, ExPASy or Brenda.

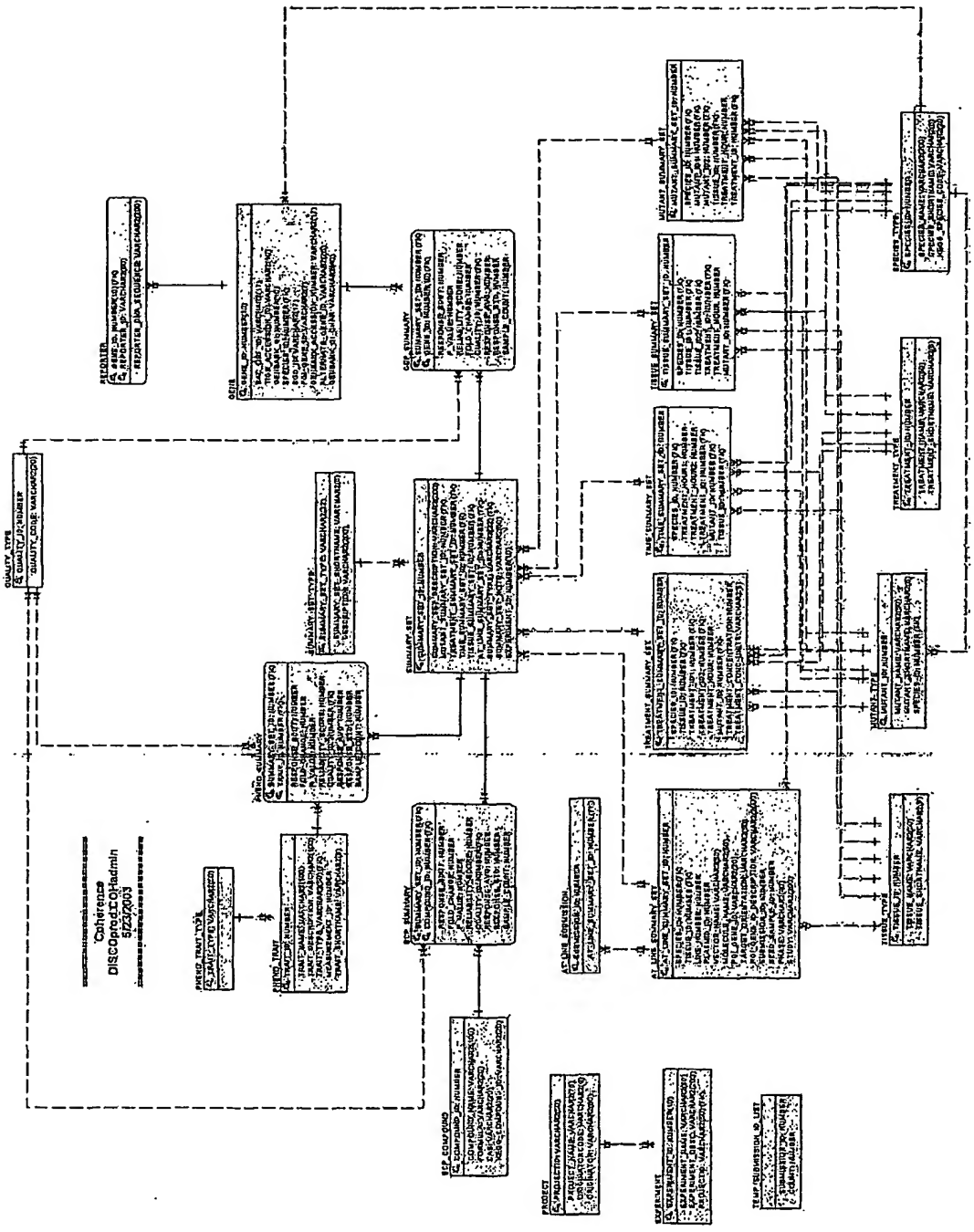

Figure 2